PATENT ABSTRACTS OF JAPAN

(11)Publication number : 09-062693
(43)Date of publication of application : 07.03.1997

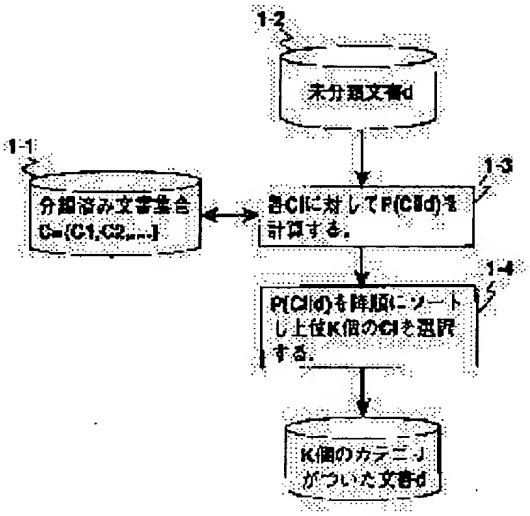
(51)Int.Cl. G06F 17/30

(21)Application number : 07-215670
(22)Date of filing : 24.08.1995

(71)Applicant : HITACHI LTD
(72)Inventor : IWAYAMA MAKOTO
MOTODA HIROSHI

(54) DOCUMENT CLASSIFYING METHOD ACCORDING TO PROBABILITY MODEL

(57)Abstract:
PROBLEM TO BE SOLVED: To automatically classify a document without interpolating any data even when features are applied by a keyword not existent in data for exercise by expressing the document in the set of words and applying the features to the document corresponding to the event of probability for extracting certain words from that word set at random.
SOLUTION: A non-classified document 1-2 is classified by using an already classified document set 1-1. In this case, the classified document set 1-1 is expressed in C={C1, C2...}. At a processing function 1-3, the 'probability P (Ci|d) for a non-classified document (d) to be classified in a category Ci' is calculated. Namely, the 'probability P (Ci|d) for the non-classified document (d) to be contained in the document set already classified as Ci' is calculated. At a processing function 1-4, the P (Ci|d) calculated by the processing function 1-3 is sorted in the descending order and K pieces of high-order categories are selected, for example, and defined as categories provided in the document (d).



LEGAL STATUS

[Date of request for examination] 02.11.2000

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

(19)日本国特許庁（ＪＰ）　(12) **公開特許公報（Ａ）**　(11)特許出願公開番号

特開平9－62693

(43)公開日　平成９年(1997)３月７日

(51)Int.Cl.[6]	識別記号	庁内整理番号	Ｆ Ｉ		技術表示箇所
Ｇ０６Ｆ　17/30			Ｇ０６Ｆ　15/401	３１０Ａ	
				３１０Ｄ	

審査請求　未請求　請求項の数１　ＯＬ　（全　３　頁）

(21)出願番号　特願平7－215670

(22)出願日　平成７年(1995)８月24日

(71)出願人　000005108
株式会社日立製作所
東京都千代田区神田駿河台四丁目６番地

(72)発明者　岩山　真
埼玉県比企郡鳩山町赤沼2520番地　株式会社日立製作所基礎研究所内

(72)発明者　元田　浩
埼玉県比企郡鳩山町赤沼2520番地　株式会社日立製作所基礎研究所内

(74)代理人　弁理士　小川　勝男

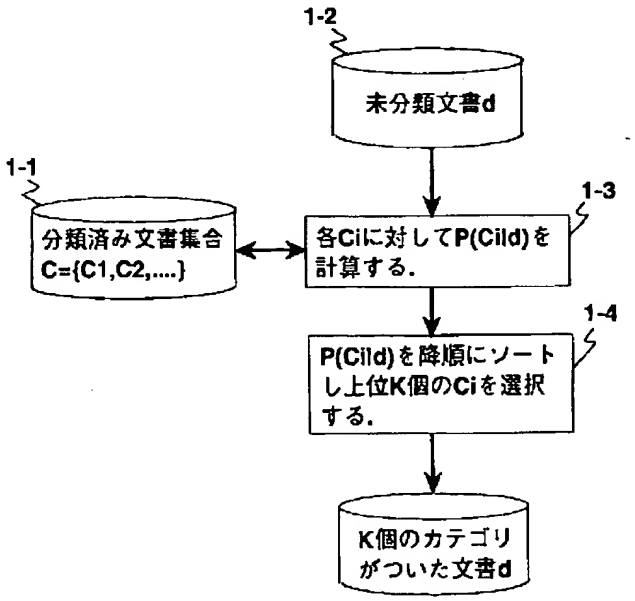
(54)【発明の名称】　確率モデルによる文書分類方法

(57)【要約】

【目的】　分類済みの文書集合（訓練用データ）をもとに新たな文書を分類する際、訓練用データが不十分な場合でもデータ補完を行うことなしに文書分類が可能になる文書の確率的特徴付けの方法とそれに基づく文書分類の方法の提供。

【構成】　特徴付けの対象となる文書集合Ｃは、それが含む単語の集合Ｗで表現される。ここで、単語集合Ｗから単語を無作為抽出する事象を考え、無作為抽出した単語がある特定の単語ｗｉと等しいという事象をＴ＝ｗｉとおき、確率Ｐ（Ｔ＝ｗｉ｜Ｃ）を推定する。事象Ｔ＝ｗｉは全てのｗｉに関して背反であるため、Ｗに含まれる全ての単語ｗｉに対して確率Ｐ（Ｔ＝ｗｉ｜Ｃ）を推定し、それらを総和した確率は、文書集合Ｃを単語集合Ｗで特徴付けたことになる。

図１



【特許請求の範囲】
【請求項１】訓練用データとしての分類済みの文書をもとに新たな文書を確率的に分類する際、文書を単語の集合として表現しておき、その集合から単語をランダムに抽出する確率事象により文書を特徴付ける方法。
【発明の詳細な説明】
【０００１】
【産業上の利用分野】本発明は、大量の文書を確率的に分類する手法に係わるものであり、従来は人手で行っていた文書分類を自動的に行う仕組みを与えるものである。
【０００２】
【従来の技術】確率的に文書分類を行うためには、文書をいかに特徴付けるかが問題になる。従来の手法は、「文書がある単語（キーワード）でインデックスされる／されない（文書にある単語が含まれる／含まれない）」という基本事象により文書の特徴付けを行っていた。例えば、文書の集合Ｃを単語ｗで特徴付けることを考えると、「文書集合Ｃの中からランダムに抽出した文書が単語ｗでインデックスされる確率」をＰ（ｗ＝１｜Ｃ）と書き、この確率により文書集合Ｃの特徴付けを行っていた。ここで、Ｐ（ｗ＝１｜Ｃ）はＣの中のｋ個の文書が単語ｗを含んでいる場合、「ｋ／（Ｃに含まれる文書数）」で推定できる。複数の単語集合Ｗ＝｛ｗ１、ｗ２、．．．、ｗｍ｝で文書集合Ｃを特徴付けるには、
P(w1=1|C)*P(w2=1|C)*. . . *P(wm=1|C)　　(1)
を計算すればよい。
【０００３】文書分類では、分類に先だって分類済みの文書があらかじめ訓練用データとして与えられている。今、文書集合Ｃとして、訓練用データの中で同じカテゴリｃに分類されている文書集合を設定すれば、前記(1)式の確率はカテゴリｃを特徴付けることに相当し、この特徴付けを用いて新たな文書を分類することが可能になる。この際、これから分類しようとする文書ｄも同じく単語で特徴付けるのだが、これは前記(1)式においてＣ＝｛ｄ｝とした場合に相当する。
【０００４】
【発明が解決しようとする課題】上記従来技術の問題点は、もしＣの中の文書どれもがある単語ｗｉを持っていなければＰ（ｗｉ＝１｜Ｃ）が０になるため(1)の確率全体も０になってしまうことである。このような単語は文書の特徴付けとして使わないのが望ましいが、事前に単語ｗｉを特定することは難しい。そこで従来技術では、データ補完（スムージング）の手法によりＰ（ｗｉ＝１｜Ｃ）が０にならないような補正を行う。ところが、補完の正当性を保証することは一般に困難である。
【０００５】本発明の目的は、上記の状況において、文書集合Ｃを特徴付ける全体の確率が上記のような単語ｗｉに影響されないような頑強な文書特徴付けの確率的方法を提供することである。
【０００６】
【課題を解決するための手段】上記目的は、文書を単語の集合で表現し、単語集合からある単語をランダムに抽出するという確率事象で文書を特徴付けることで達成される。
【０００７】
【作用】文書集合Ｃ＝｛ｄ１、ｄ２、．．．、ｄｎ｝を単語集合Ｗ＝｛ｗ１、ｗ２、．．．ｗｍ｝で特徴付けることを考える。ここで、Ｃの各文書ｄｉは文書ｄｉに含まれるキーワードの集合で表現されている。例えば、文書ｄｉが単語ｗ１を１個、ｗ５を３個、ｗ８を２個含んでいれば、
di = {w1、 w5、 w5、 w5、 w8、 w8}　　(2)
となる。Ｃの表現形式は、Ｃに含まれる各々の文書表現形式（つまり上記の各集合ｄｉ）の和集合となる。
【０００８】C = d1 v d2 v . . . v dn
（“v”は和集合の演算子）　　(3)
ここで、単語集合からある単語を無作為抽出する事象を考え、無作為抽出した単語がｗｉと等しいという事象をＴ＝ｗｉとおく。この事象は全てのｗｉに関して背反であるため、各事象に対して総和した確率、
P(T=w1|C)+P(T=w2|C)+. . . +P(T=wm|C)　　(4)
は、文書集合Ｃを単語集合Ｗで特徴付けたことに相当する。Ｐ（Ｔ＝ｗｉ｜Ｃ）は、Ｃにｗｉがｋ個含まれている場合、「ｋ／（Ｃに含まれる単語数）］で推定できる。(4)式において、全体の確率が各単語に関する確率の和の形になっていることに注意されたい。
【０００９】従来の確率(1)は積の形になっている。よって、Ｃのどの文書も単語ｗｉを持っていない場合を考えると、Ｐ（Ｔ＝ｗｉ｜Ｃ）は０になるが、和形式としたことにより全体の確率(4)も０になってしまうことはない。従来の確率(1)では積形のため全体の確率が０になってしまうことが問題であった。
【００１０】
【実施例】以下、本発明の実施例である自動文書分類について説明する。
【００１１】図１に自動文書分類の概要を示す。自動文書分類では、既に分類済みの文書集合１－１を用いて、未分類の文書１－２を分類する。ここで、分類済み文書集合１－１は、Ｃ＝｛Ｃ１、Ｃ２、．．．｝と表現される。各Ｃｉはカテゴリ Ｃｉと分類されている文書の集合である。よって、例えば、ある文書がカテゴリＣ１とカテゴリＣ３に分類されている場合、その文書は文書集合Ｃ１、Ｃ３両方に含まれることになる。
【００１２】処理機能１－３においては、「未分類の文書ｄがカテゴリＣｉに分類される確率Ｐ（Ｃｉ｜ｄ）」を計算する。言い替えると、「未分類の文書ｄが、既に“Ｃｉ”として分類されている文書集合に含まれる確率Ｐ（Ｃｉ｜ｄ）」を計算することになる。
【００１３】この確率Ｐ（Ｃｉ｜ｄ）を計算するため

に、本発明で提案した文書の特徴付けを用いる。具体的には、「ある単語集合から無作意に抽出したキーワードがｗｊと等しい」という事象Ｔ＝ｗｊを考える。この事象は、全ての単語に関し背反であるため、各事象についてＰ（Ｃｉ｜ｄ）を条件付けると、

$$P(C|d) = P(C) * sum\_\{j\} [P(T=wj|C)*P(T=wj|d)/P(T=wj)] \quad (6)$$

となる。(6)式の

sum_{j} [P(T=wj|C)*P(T=wj|d)]

は、文書ｄとカテゴリＣｉを単語集合で同時に特徴付けたことに相当している。

【００１４】各確率は以下の方法で推定できる。

【００１５】Ｐ（Ｔ＝ｗｊ｜Ｃｉ）＝「Ｃｉに含まれる単語ｗｊの数／Ｃｉに含まれる単語数」

Ｐ（Ｔ＝ｗｊ｜ｄ）＝「ｄに含まれる単語ｗｊの数／ｄに含まれる単語数」

Ｐ（Ｔ＝ｗｊ）＝「全文書に含まれる単語ｗｉの数／全文書に含まれる単語数」

Ｐ（Ｃｉ）＝「Ｃｉに含まれる文書数／全文書数」

上記(6)式を用いると、各候補カテゴリＣｉについてＰ（Ｃｉ｜ｄ）が計算出来る。処理機能１－４において、

$$P(Ci|d) = sum\_\{j\} [P(Ci|T=wj)*P(T=wj|d)] \quad (5)$$

(sum_{j}は全てのjに対する総和)

となる。ここで、ベイズの定理を用いてＰ（Ｃｉ｜Ｔ＝ｗｊ）を書きかえると、

計算したＰ（Ｃｉ｜ｄ）を降順にソートして、例えば、上位Ｋ個のカテゴリを選択し、文書ｄが持つカテゴリとする。

【００１６】

【発明の効果】本発明で提案する文書の特徴付けによると、訓練用データの中に存在しないキーワードで特徴付けを行っても、データの補完をすることなしに文書自動分類が可能になる。

【図面の簡単な説明】

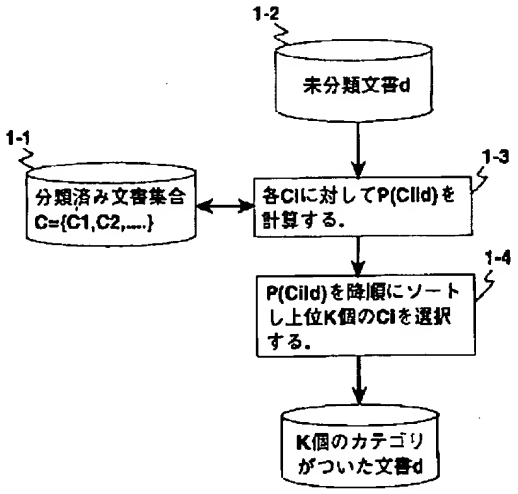
【図１】本発明の実施例の処理手順を示した図。

【符号の説明】

１－１：既に分類済みの文書集合、１－２：未分類の文書

【図１】

図1



PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2000-155758
(43)Date of publication of application : 06.06.2000

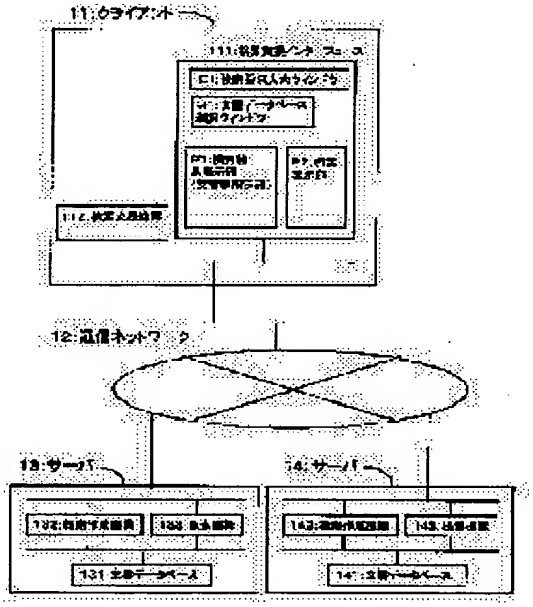
(51)Int.Cl. G06F 17/30

(21)Application number : 10-328940
(22)Date of filing : 19.11.1998

(71)Applicant : HITACHI LTD
(72)Inventor : IWAYAMA MAKOTO
NISHIOKA SHINGO
NIWA YOSHIKI
TAKANO AKIHIKO

(54) METHOD AND SERVICE FOR DOCUMENT RETRIEVAL FROM PLURAL DOCUMENT DATA BASES

(57)Abstract:
PROBLEM TO BE SOLVED: To efficiently realize document retrieval for checking the correlations between document data bases.
SOLUTION: The document data bases 131 and 141 are provided with summary generating mechanisms 132 and 142 and retrieving mechanisms 133 and 143 and are connected as servers 13 and 14 to a communication network 12. A client 11 obtains a relevant document group among document groups in a specified document data base through its summary. The obtained summary is sent to the other server to perform retrieval corresponding to the transferred summary from the document data base on the server at the transfer destination.



LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]